IPCC AR5 data management as seen from GFDL (today)

AR5 Data Management Meeting
Princeton NJ

V. Balaji

Princeton University

16 October 2007

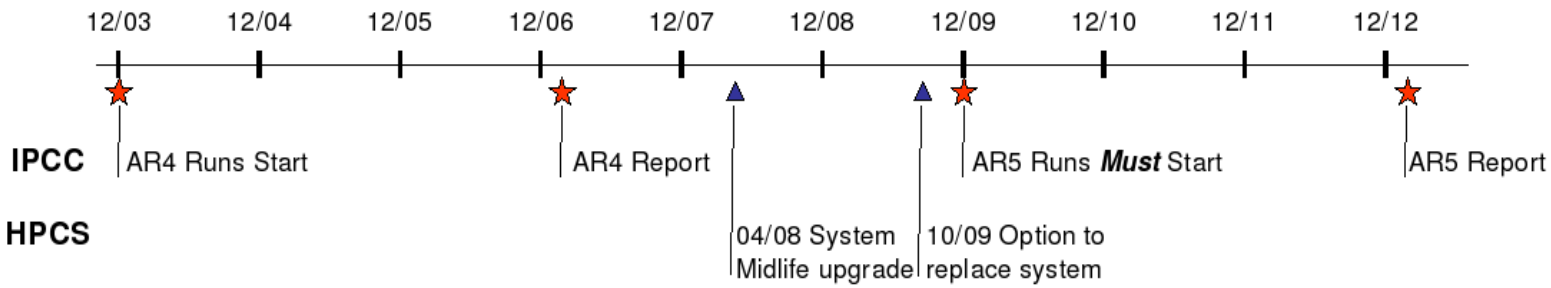
Talk outline. . .

1. What will be run and when
2. The data pipeline
3. Issues for AR5

Talk outline . . .

1. What will be run and when
2. The data pipeline
3. Issues for AR5

Projected (guess) AR5 timeline



- AR5 Cycle
  - Report issued Feb 2013 (let's assume. . .)
  - . . . then runs must start December 2009 at the latest
  - Potentially 2 streams:

    traditional century-scale, control, $2\times CO_2$, historical, scenarios, carbon cycle (new)

    "short-term" initialized decade-scale ensemble projections (out to 2030 or 2050)
- GFDL issues
  - System upgrade and model development cycles overlap. . .
  - Significant resource impact on Modeling Systems and Technical Systems

The models (guess)

We're not in a position to predict data volumes yet. This lists some candidate models, with implications for AR5 data management.

ESM2.1 Earth System Model including carbon cycle, dynamic vegetation, historical land surface forcings, ocean biogeochemistry: resolutions similar to CM2.1 but many new fields ("CMOR tables");

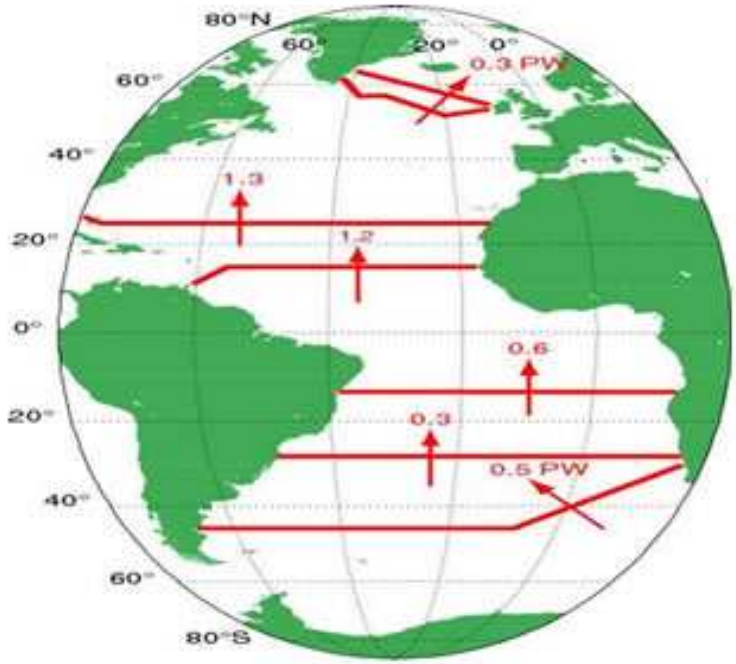
CM2.4 physical climate model, increased resolution (16X data volume);

FVCS atmospheric models may use the cube-sphere dynamical core, requiring use of the mosaic gridspec: would benefit from server-side regridding capabilities, as we may not be able to pre-compute all possible fields on standard grids.

Surprises last time we budgeted only for CM2.0, then CM2.1 came along. . . we still have two ocean models in play late in the game this time.

Horizontal regridding: poleward heat transport

- Atmospheric data:
  - v, T, q, $\overline{v'T'}$, $\overline{v'q'}$
  - $F^{\uparrow}_{\text{sfc}}$, $F^{\uparrow}_{\text{TOA}}$
  - p_s
- Ocean data:
  - v, T, $\overline{v'T'}_{\text{total,gyre,eddy,...}}$: total and per basin.
  - meridional mass overturning circulation: total and per basin

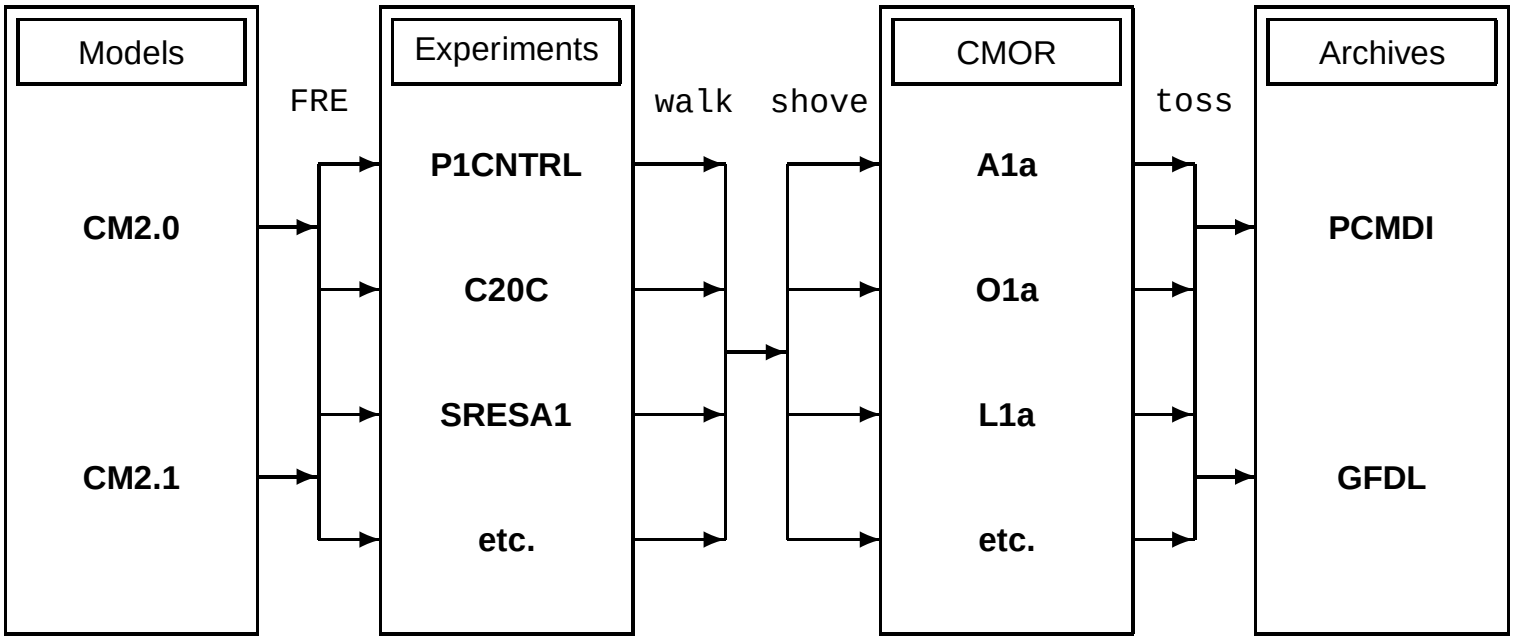


`http://www-pcmdi.llnl.gov/ipcc/project_detail.php?ipcc_-subproject_id=174`

Talk outline . . .

1. What will be run and when
2. The data pipeline
3. Issues for AR5

The GFDL data pipeline, AR4 vintage



- time- and data-intensive;
- multiple access episodes for the same datasets;
- would be ideal if FRE already produced compliant data.

The GFDL data pipedream

- FMS I/O already produces CF-compliant data
- FRE enters experiments directly into GFDL Curator DB.
- Curator DB applies metadata transformations as specified by modeling campaign (IPCC, TFSP, CFMIP, O_3 . . .); perhaps this is done in NcML and is low-cost? I'm wary of this approach for two reasons:
  - Bulk of our data transfer volumes is still in `ftp`/`wget`;
  - non-standard short names.
- Curator DB has interface to metadata harvester.

Talk outline . . .

1. What will be run and when
2. The data pipeline
3. Issues for AR5

Issues for AR5

- Native grid data (Curator/Metafor for spec; originating site takes responsibility for regridding algorithm; who deploys it as a web service?);
- Increased use of forcing fields and initial condition fields (Curator/Metafor for spec; CMIP4 for content?);
- Are the actual stored “naked” files going to be useless without metadata or data transformations?
- Is the system for exchanging metadata going to be ready in time?
- Is the system for exchanging metadata going to be fault-tolerant? Who is responsible for failures to hand off?
- Who is responsible for the software stack at the server node?